caBIG™
cancer Biomedical Informatics Grid ™

Center caBIG™ Goals Introduction

R. Mark Adams, Ph.D.

Agenda

- Brief Recap of the Center caBIG™ Goals Document
- Discussion of Questions Submitted in Advance
- Open Question and Answer Session

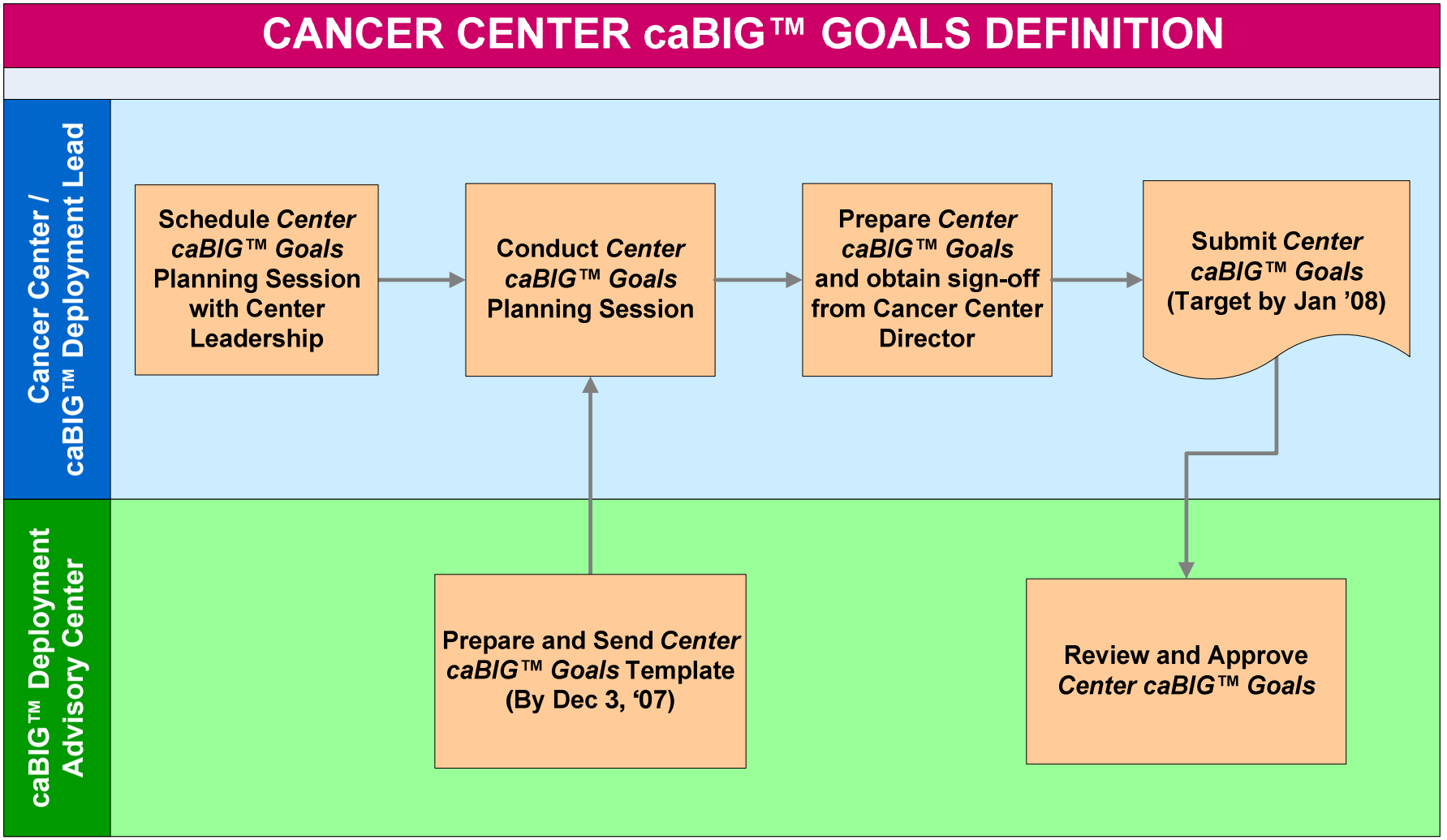
Center caBIG™ Goals Definition

- The What, The When & The How
- Center caBIG™ Goals Document
- Center caBIG™ Capability Analysis Report Integration
- Center caBIG™ Goals Definition Planning Activities

Center caBIG™ Goals Definition – The What?

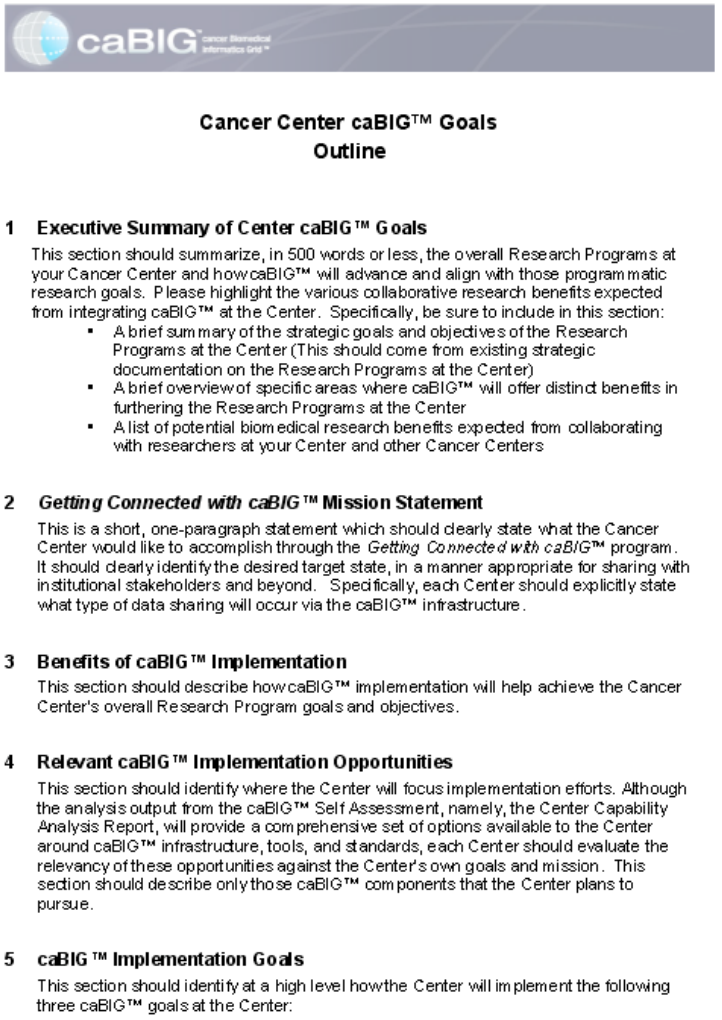
Center caBIG™ Goals Definition	Description
Summarize the Research Programs at your Center	• Summarize the strategic goals and objectives of the Research Programs at your Center
Leverage caBIG™ to advance the Research Program at your Center	• Align Center caBIG™ goals with overall Center Research Programs • Highlight specific benefits offered by implementing caBIG™ • Identify relevant caBIG™ implementation opportunities • Identify specific goals for implementing caBIG™ infrastructure and tools (Example, Manage all biospecimens using caTissue within 12 months; Submit Summary 3/ Summary 4 reports electronically; etc.) • Develop a plan for managing Center capabilities and potential challenges in implementing caBIG™
Highlight collaborative research benefits from integrating caBIG™ at your Center	• List potential research benefits expected from collaborating with researchers at your Center and other Cancer Centers

Center caBIG™ Goals Definition – The When & The How?



Center caBIG™ Goals Document

1. Executive Summary of Center caBIG™ Goals
2. *Getting Connected with caBIG™* Mission Statement
3. Benefits of caBIG™ Implementation
4. Relevant caBIG™ Implementation Opportunities
5. caBIG™ Implementation Goals
6. Managing Capabilities and Potential Challenges for caBIG™ Implementation



caBIG

Cancer Center caBIG™ Goals Outline

1 Executive Summary of Center caBIG™ Goals

This section should summarize, in 500 words or less, the overall Research Programs at your Cancer Center and how caBIG™ will advance and align with those programmatic research goals. Please highlight the various collaborative research benefits expected from integrating caBIG™ at the Center. Specifically, be sure to include in this section:

- A brief summary of the strategic goals and objectives of the Research Programs at the Center (This should come from existing strategic documentation on the Research Programs at the Center)
- A brief overview of specific areas where caBIG™ will offer distinct benefits in furthering the Research Programs at the Center
- A list of potential biomedical research benefits expected from collaborating with researchers at your Center and other Cancer Centers

2 *Getting Connected with caBIG™* Mission Statement

This is a short, one-paragraph statement which should clearly state what the Cancer Center would like to accomplish through the *Getting Connected with caBIG™* program. It should clearly identify the desired target state, in a manner appropriate for sharing with institutional stakeholders and beyond. Specifically, each Center should explicitly state what type of data sharing will occur via the caBIG™ infrastructure.

3 Benefits of caBIG™ Implementation

This section should describe how caBIG™ implementation will help achieve the Cancer Center's overall Research Program goals and objectives.

4 Relevant caBIG™ Implementation Opportunities

This section should identify where the Center will focus implementation efforts. Although the analysis output from the caBIG™ Self Assessment, namely, the Center Capability Analysis Report, will provide a comprehensive set of options available to the Center around caBIG™ infrastructure, tools, and standards, each Center should evaluate the relevancy of these opportunities against the Center's own goals and mission. This section should describe only those caBIG™ components that the Center plans to pursue.

5 caBIG™ Implementation Goals

This section should identify at a high level how the Center will implement the following three caBIG™ goals at the Center:

caBIG

Center caBIG™ Capability Analysis Report Integration

- Summarizes institutional and infrastructure readiness
- Identifies technology capabilities
- Assesses staff and investigators' capabilities
- Indicates which tools to review for adoption/adaptation
- Identifies current data sharing capabilities
- Provides recommendations for Center to implement caBIG™ infrastructure and tools
- Identifies challenges to implementation

Center caBIG™ Goals Planning Activities

1. Review Center caBIG™ Goals Template posted to:

   Center Deployment Listserv at
   https://list.nih.gov/archives/center_deployment.html

   or Center Deployment Leads Web Page at
   https://cabig.nci.nih.gov/center_deployment

2. Identify Center Leadership/ Stakeholders and Schedule Planning Sessions

3. Review caBIG™ Deployment Process with your Center Leadership/ Stakeholders

4. Define Center caBIG™ Goals with Center Leadership/ Stakeholders

5. Draft Center caBIG™ Goals Document

caBIG™
cancer Biomedical Informatics Grid ™

Discussion of Questions Submitted in Advance

Data Sharing Approach

One of the goals of the current grant is to share data using a caGrid node. What is the simplest method of doing this?

- The easiest way to share data on the caGrid is through one of the Grid-enabled applications from the application bundles. You will need to install a local caGrid node at your local institution, and then make use of specific caGrid-enabled applications. By loading local data into the local application, the data can then be made available on the Grid.

caBIG™ and caGrid Components

In addition to installing a node, what other components need to be installed? For example, what is required to share microarray data? What is required to share clinical data?

- Installing a caGrid node creates local infrastructure required to allow specific Grid Services to be placed at your site
- In order to share data with the wider community, you would need to install additional applications from the bundles that would do the actual hosting of data and caGrid Services.
- For microarray data, you could use the caArray product (watch for the upcoming demo for more details on how this works!)
- For clinical data, a tool like CTODS (Clinical Trials Object Data System) can be used to aggregate and host specific data for communication on the grid
- The standard caGrid security infrastructure can then be used to regulate access to users and groups who are allowed to view the specific data sets
- Note that it is possible to install all the caGrid infrastructure for purely *local* data sharing with sensitive data such as identified patient information, but that is more complex and requires more effort than a standard caGrid installation
- The applications vary with respect to the level of difficulty for installing and testing.

General Costs/Time for Adapting

What is the general timeline and costs for adapting pre-existing in-house software?

- This will vary significantly based on your own institutions existing infrastructure. By comparing the specific requirements for the caGrid components to what your own institution already supports as a standard part of running its business, you can get an estimation of the specific costs.
- The deployment support center can also help locate and identify any additional information you would need to develop these estimates, and put you in contact with those who can supply specific needed information
- We are working on a set of “design patterns” that should serve to provide some general guidance in helping each institution come up with implementation plans that reflect the unique nature of their respective sites.
- There are a range of existing processes and models within the program that you can leverage to provide a framework for your own local efforts.

Adapting Specific Systems

How, in general, can existing local and/or commercial systems such as Oncore and Epic be used in conjunction with caBIG™?

- In general, the answer will be highly dependent on your institution's specific tooling, capabilities, interests, and business/vendor relationships
- We can share some general illustrations, though to provide examples of how such adaptation can be achieved:
- For a center that has clinical data to share, a relatively straightforward approach can be to develop extract, transform and load (ETL) scripts from an existing clinical data store, map that data to the caBIG-standard data elements, and place the data in a caBIG-compatible data store such as CTODS.
- Similarly, a tool such as caXchange could be used to take industry-standard data streams in the form of messages (or even ASCII tables) and transform the information into caBIG compatible data which can then be placed in a caBIG compatible data repository such as CTODS to provide access
- Many other possibilities exist, such as developing caBIG compatible application programming interfaces (APIs) on top of existing systems, but this requires specific capabilities or experience in the institution or its subcontractors.

Adaptation Redux

Does adaptation, for example, mean using caCore to convert a system into one that is caBIG compatible (which involves registering concepts and entities with the caDSR, etc.)?

- Adaptation can take many different forms, depending on the skills, requirements and needs of the participating institution.
- In general, adaptation just means that the existing systems are provided a means to place designated data where it can be shared using caBIG standards and infrastructure
- This can include using caCORE and other caBIG tools to directly provide caBIG compatible interfaces on existing tools. The tools exist, and have been tested and used throughout the program by a variety of different institutions, but they can require effort and specific experience which may not exist at all centers.
- A possible solution for this is to work with service providers that have experience with the tools, and can help get your institution up to speed with the systems quickly.

caBIG™ Installation Costs

What is the cost of installation for a particular caBIG component. For example, what is the cost for installation for a caGrid node?

- Specs for a machine which will ONLY host caGrid core services *and* a single caGrid app at a time that is *not* graphics intensive, and which will *not* host the database it needs to access, are:
- Intel (or equivalent from AMD) Dual Core CPU at 2 ~ 3 GHz clock speed, 4 GB of RAM, 2 disks at 100 - 150 GB capacity each for 200 - 300 GB total storage. (Again, contingent on if the database can be hosted on a separate machine from this one.)
- If the grid app is graphics or computationally intensive, then it would be a good idea to double the RAM to 8 GB and increase the processor to perhaps a quad core at 2 ~ 3 GHz.
- Need staff with more of a computer science / SW Engineering background and experience that simply a bioinformatics background. We are currently dealing with an individual trying to install caGrid that is calling us for help with simple windows commands.
- Historically some difficulties with highly-outsourced IT departments where they have no local control over their own environment.
- If you have all the necessary people together to provide all the security access, configure the system, etc., I would allow 2-2.5 days for installing and testing the caGrid system.
- With a team onsite, and reasonable access/control of systems an outside team could complete all the necessary activities in a week or two
- Worst case, where it is all done via the phone with no remote access, it can take more than a month

More Hardware Information

For more Hardware-intensive, or larger-scale installations, here are some rules of thumb:

- While the database may be installed on a separate machine, we may want to install caGrid and IVI Middleware on the same machine as the DB for performance and possibly API access reasons.
- For image analysis, a lot of research algorithms we've come across are not really multi threaded. The number of cores will help to support simultaneous users but not necessarily in the speed of processing in those cases. (this is to set the expectations properly.) Quad core is not that much more expense though.
- Hardware RAID 5. This is the best option for performance, redundancy, and future expansion.
- If performance is critical go for fast 10K RPM disks SATA or SCSI, otherwise 500GB disks are at $150 to $200 range for 7200 RPM SATA.
- If there is anticipated need to support graphics based processing, then the machine should be in 3U or larger rack mount case or in desktop/tower form factor for adding double width graphics hardware.
- Redundant power supply if this is a critical server. Battery backup.
- Gigabit Ethernet, two if possible on the motherboard.
- 8GB is fine. Rule of thumb is 2GB per core.

Hardware Upgrade Cycle

What is the hardware upgrade cycle? Our standard for servers is 3 years. What is the experience of early adopters in this area?

- Given the limited amount of time we have had to run this activity, we have not yet collected enough data to be able to answer this question with certainty, but we suspect that the standard business metrics used by your institution would be sufficient
- What may differ in any given installation has to do with the use to which the caBIG™ infrastructure is being put. An installation which is serving quite well for largely text-based information from a biospecimen management system might experience significant performance degradation if it began to be used for large amounts of biomedical imaging data, for example.
- Planning should take into account both the current and likely future scenarios in order to ensure satisfactory performance

Costs for Managing a Grid Node

What is the cost of maintaining a caGrid node? What type of FTE and % effort is required? I understand that this will be institution dependent, but some range would be useful.

- Once the system is installed and running, maintenance is minimal except for upgrades and if someone messes up the configuration (which is known to have happened). The base line configuration needs to be maintained, including several tools:

- caGrid Version
- Browser
- Database
- Jboss App Server
- JSDK
- Apache Ant
- Apache Axis
- Hibernate
- Common Security Module
- caCORE SDK
- caCORE
- Globus Tookit
- Tomcat
- HTTP Server

caBIG™
cancer Biomedical Informatics Grid ™

Open Question & Answer Session